Relaxing the Exponential Assumption

This Hypercard stack was prepared by:
Dennis L. Bricker,
Dept. of Industrial Engineering,
University of Iowa,
Iowa City, Iowa 52242
e-mail: dennis-bricker@uiowa.edu

author

In queues which are modeled as birth-death processes,

both the times between arrivals
and the service times

must have exponential distributions.

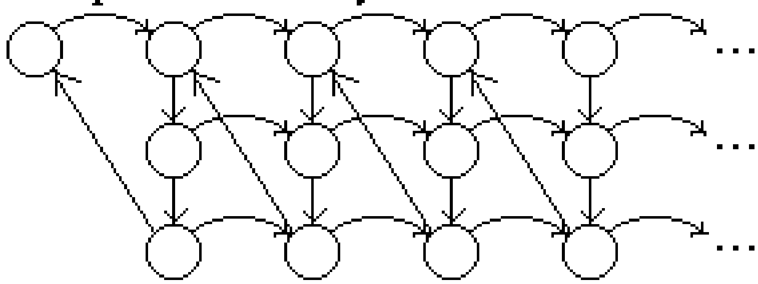
Exponential distributions have coefficient of variation equal to 1, i.e.,

$$\frac{\sqrt{\mathrm{var}[T]}}{E[T]} = 1$$

If, in an application, inter-arrival &/or service times are either more or less *regular*, what can be done?

We have seen that an $M/E_k/1$ queue, for which service times have Erlang-k distribution, can be modeled as a (continuous-time) Markov chain.

That is, the service consists of k phases, each with exponentially-distributed service time.



If service time T has Erlang-k dist'n with mean $1/k\mu$,

then

$$T = \sum_{i=1}^{k} Y_i \,,\quad E[T] = \sum_{i=1}^{k} E[Y_i] = kE[Y_i] = 1/\mu$$

$$Var[T] = \sum_{i=1}^{k} Var[Y_i] = k \times Var[Y_i] = k/(k\mu)^2 = 1/k\mu^2$$

$$\Rightarrow \text{Coefficient of variation} = 1/\sqrt{k}$$

The coefficient of variation may be made as small as we like (but >0) by increasing k.

An Erlang-k random variable is a *convolution* of k random variables with exponential dist'n, and is *more regular* than a random variable with exponential distribution.

To approximate distributions which are *less regular*, i.e., have c.v. > 1, we can use a *hyper-exponential* distribution.

$$P\{T \le t\} = F(t) = \beta\left[1-e^{-\mu_1 t}\right] + (1-\beta)\left[1-e^{-\mu_2 t}\right]$$

where $0 < \beta < 1$.

Hyper-Exponential Distribution

A hyper-exponential dist'n is a *mixture* of exponential distributions, with service rate μ defined by

$$\frac{1}{\mu} = \frac{\beta}{\mu_1} + \frac{(1-\beta)}{\mu_2}$$

and has coefficient of variation >1 and can be made arbitrarily large.

M/H/1 Queue

Arrivals are Poisson with rate λ , and service time has hyper-exponential dist'n

$$F(t) = \beta\left[1 - e^{-\mu_1 t}\right] + (1-\beta)\left[1 - e^{-\mu_2 t}\right]$$

Equivalently, suppose 2 types of customers,

Type 1 with service time distn Exp(μ_1)

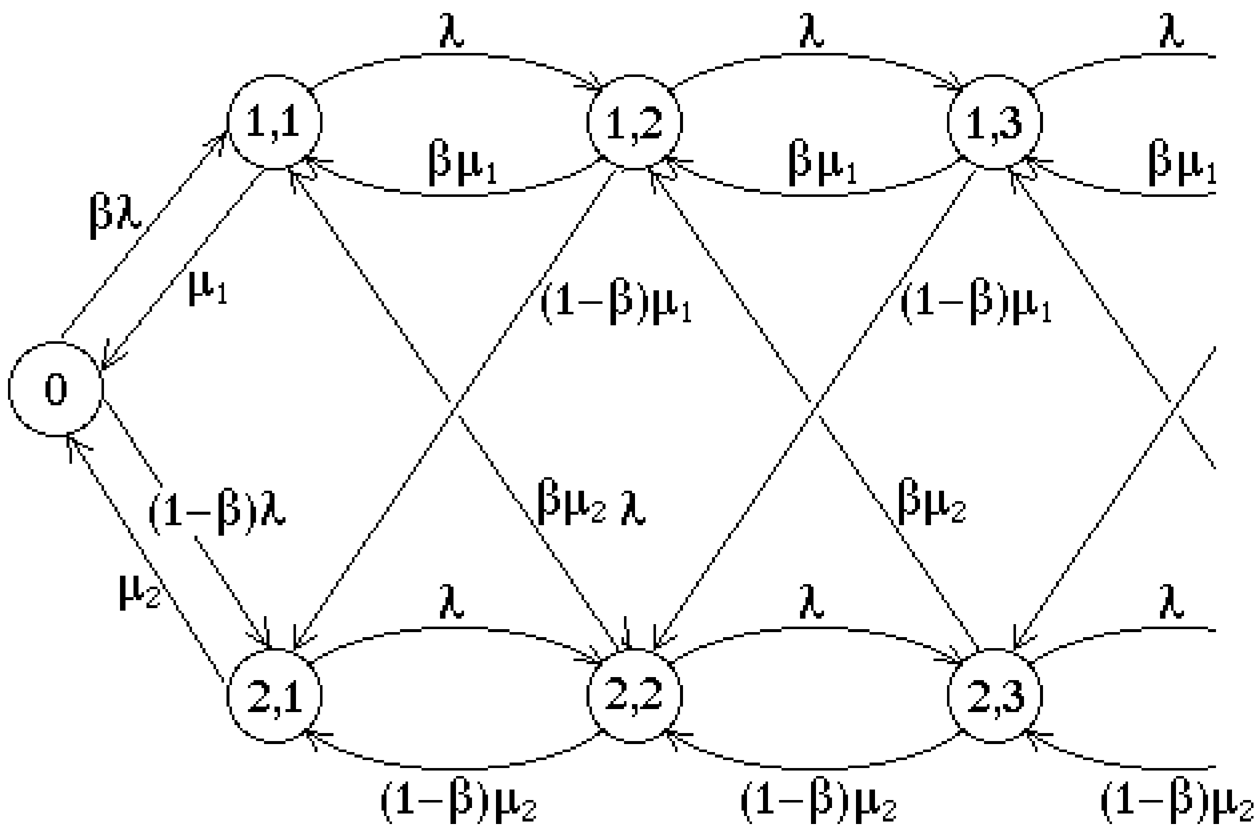
Type 2 with service time distn Exp(μ_2)

where β = fraction of customers that are type 1

States

i,j

where
- i=type of service being provided
- j=# in system



Phase-type Distribution

A distribution function H is phase-type with k phases

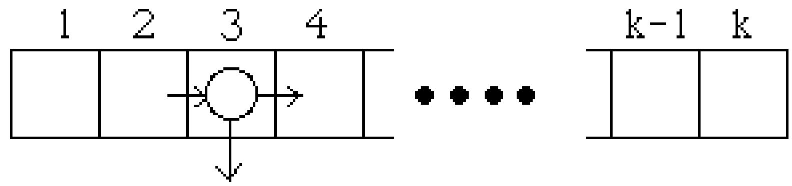
if it is $(1-\beta_1)F_1 + (1-\beta_2)\beta_1 F_1 \oplus F_2 + \cdots$
$+\beta_1\beta_2\cdots\beta_{k-1}F_1\oplus F_2\oplus\cdots\oplus F_k$

where F_i is exponential dist'n with rate μ_i

$F_1\oplus F_2$ is the convolution of F_1 and F_2,

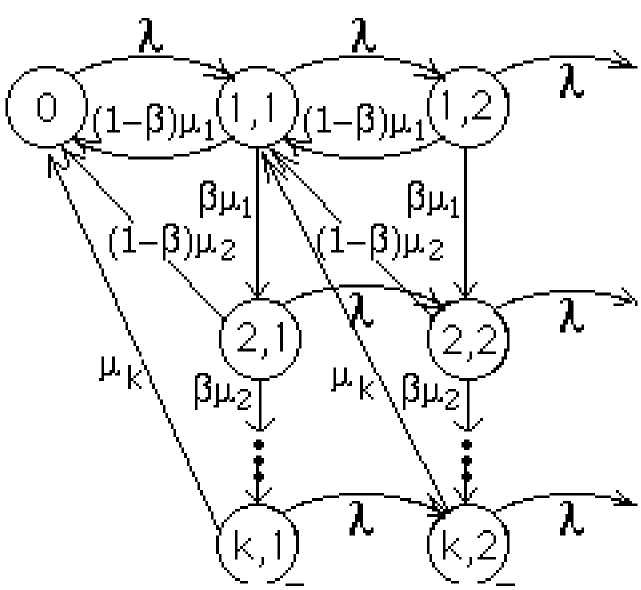
$0<\beta_i<1$ for $i=1,2,\ldots k-1$, $\beta_k=0$

Phase-type Distribution

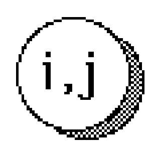


Think of the service facility as having k stages, with a service time in stage j having exponential dist'n (rate μ_j),

and upon completion of stage j, the service is complete with probability $(1-\beta_j)$ or customer enters stage j+1.



States



i=phase of service in progress
j= # of customers in the system

Phase-type Distribution

The Erlang & Hyper-exponential dist'ns are both phase-type distributions.

Any arbitrary distribution can be approximated as closely as desired by a phase-type dist'n.

example Suppose that service time T is discrete, with $P\{T=a\} = 1-\beta$ and $P\{T=b\} = \beta$, for $\beta \in (0,1)$.

That is, $T = a + (b-a) I$

where $P\{I=1\}=\beta$, $P\{I=0\}= 1-\beta$.

Consider the distribution $(1-\beta)\, E_{k'} + \beta\, E_{k'} \oplus E_{k''}$

where $E_{k'}$ is Erlang-k' with phase rate k'/a

$E_{k''}$ is Erlang-k" with phase rate k"/(b-a)

As $k' \to \infty$ & $k'' \to \infty$, this dist'n converges to that of T.

Using phase-type distributions, we can, in principle, approximate any queue as a continuous-time Markov chain....

In practice, the state space of this Markov chain may be large &/or complex, and the balance equations intractable.

As the distributions become more regular, i.e., coefficient of variation decreases, the performance measures of the queueing system usually improve.